RSAC 2025 Conference

Many Voices. One Community.

SESSION ID: SAT-T10

Securing Retrieval Pipelines in Large Language Models

Oleksandr Yaremchuk
Head of Engineering, LLM Security
Protect AI

Neal Swaelens
Head of Product, LLM Security
Protect AI

Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

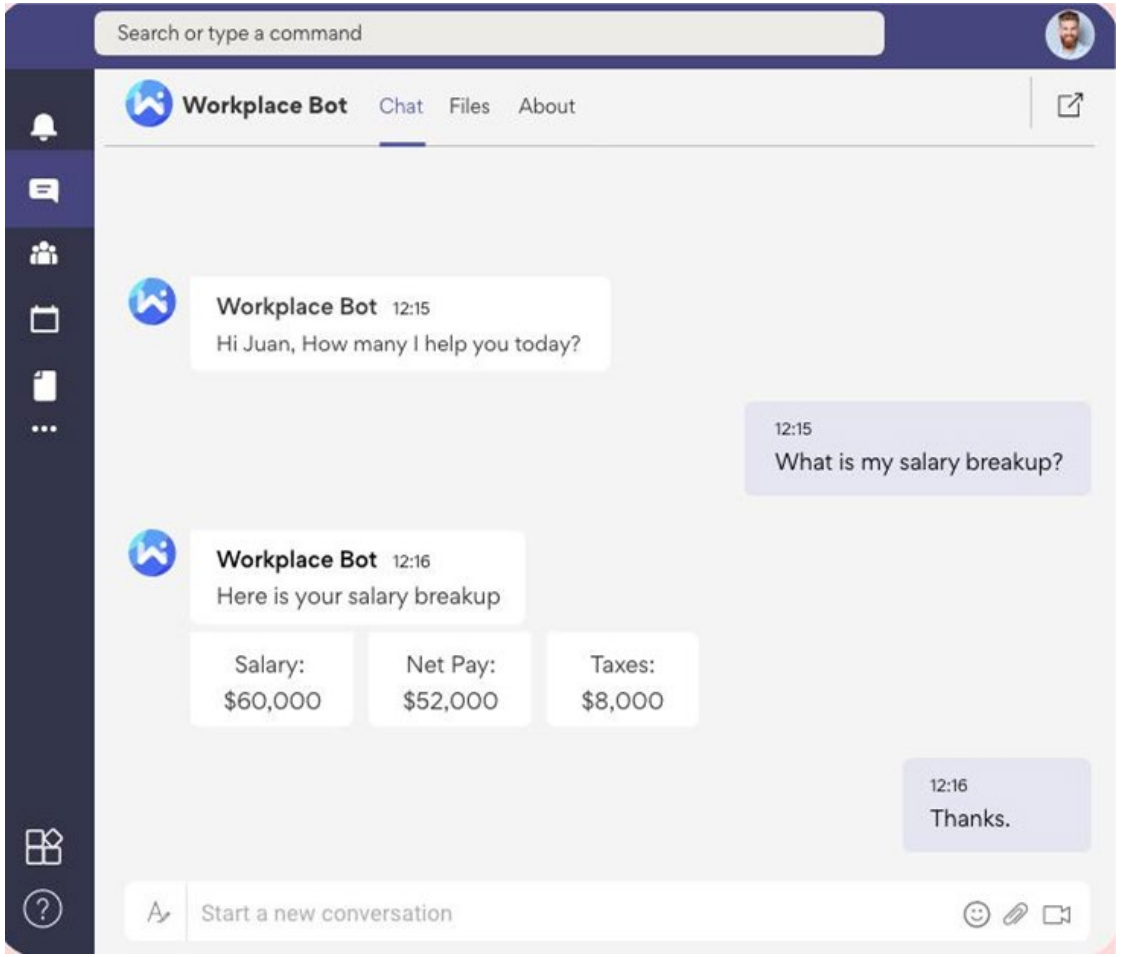
© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

This presentation is for educational purposes only and does not constitute legal, compliance, or security advice. Views expressed are those of the presenters and may not reflect Protect AI's positions. Attendees should conduct their own assessments before applying any information shared. Protect AI is not responsible for any actions taken based on this presentation.

How many of you have LLM applications running in your organization?

How many of you know whether these applications use RAG?

A Typical Enterprise LLM Application

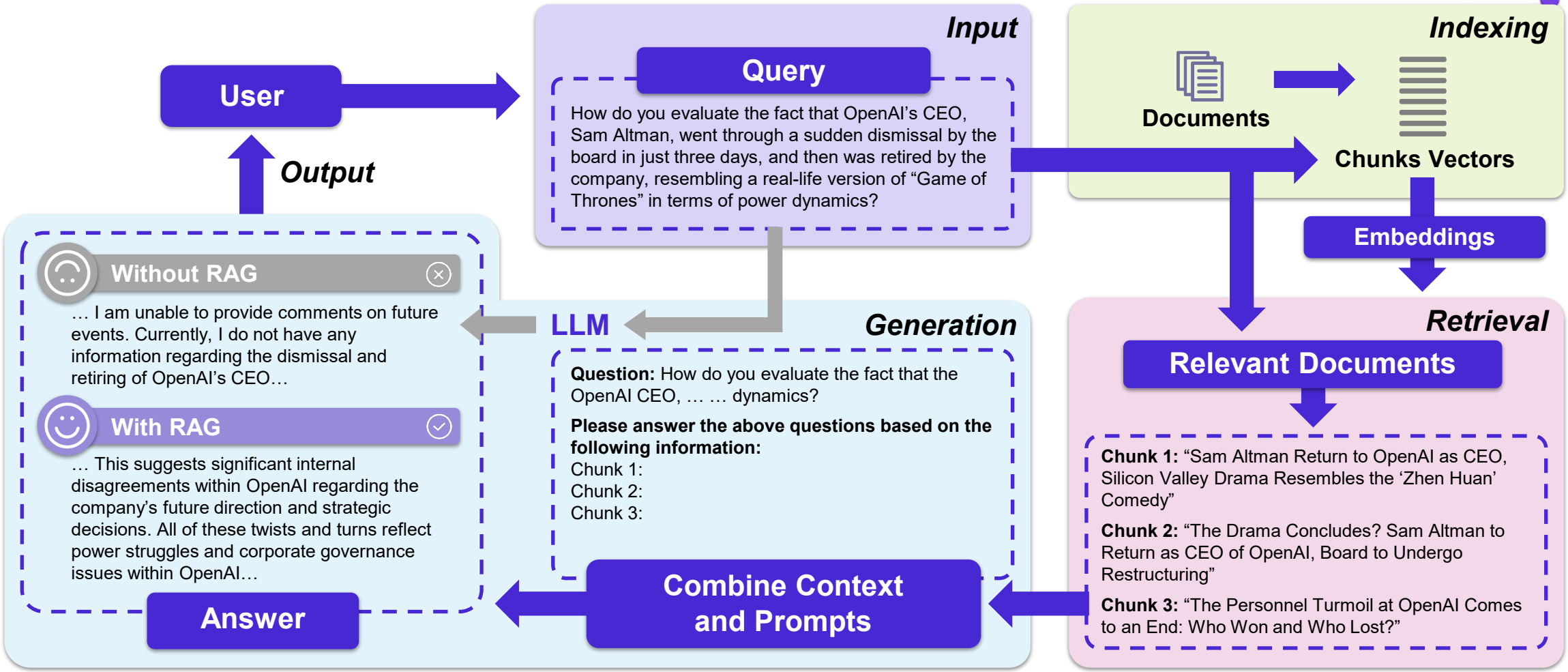


You're a customer support agent. You know this: <EXTERNAL_KNOWLEDGE>

A customer asked this: <USER QUERY>

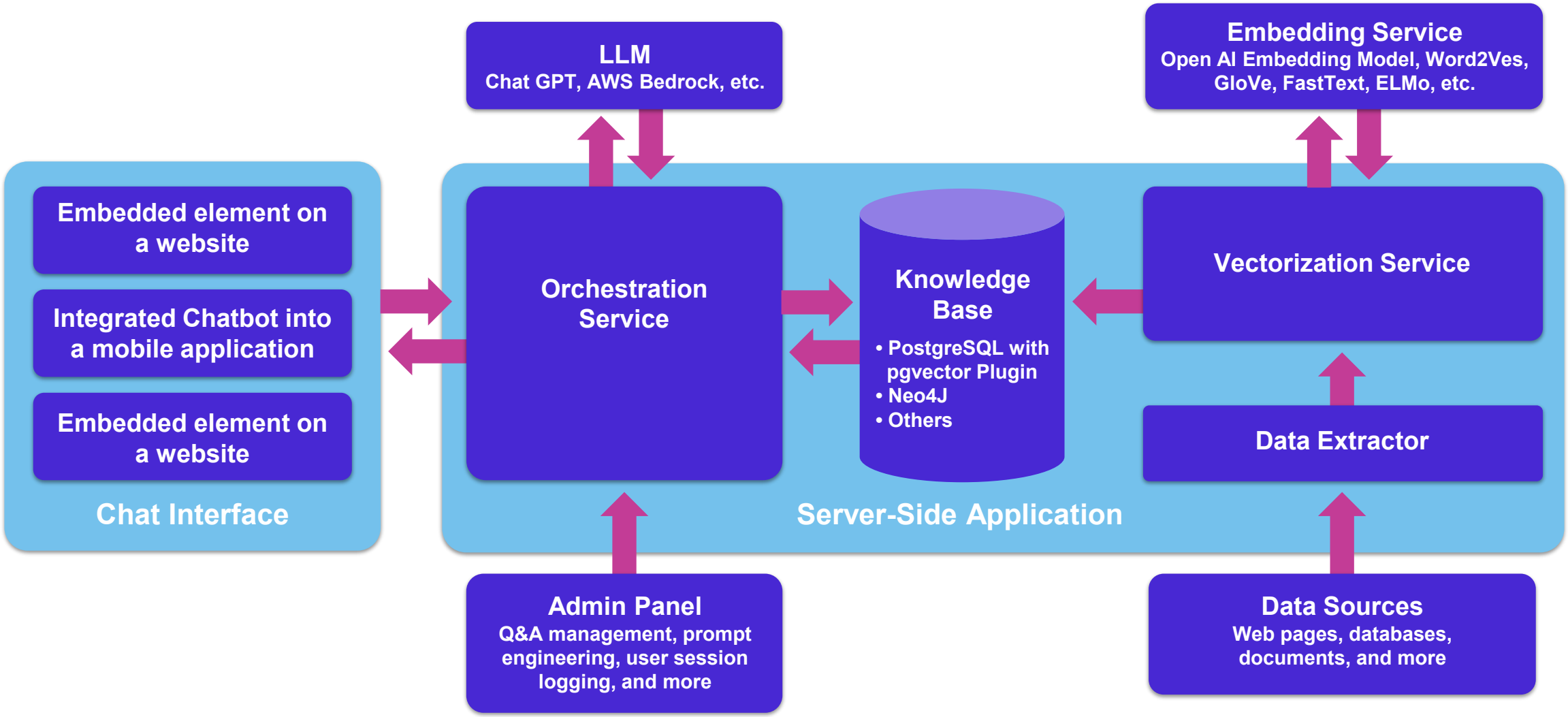
Please generate an answer for the customer.

RAG: How It Works


Input
User
Query
How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was retired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?
Indexing
Documents
Chunks Vectors
Embeddings
Output
Without RAG
… I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and retiring of OpenAI's CEO…
With RAG
… This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI…
Answer
LLM
Generation
Question: How do you evaluate the fact that the OpenAI CEO, … … dynamics?
Please answer the above questions based on the following information:
Chunk 1:
Chunk 2:
Chunk 3:
Combine Context and Prompts
Retrieval
Relevant Documents
Chunk 1: "Sam Altman Return to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"
Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"
Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"

Source: promptingguide.ai

RAG Security Will Revolve Around Securing Multiple Data Sources in Parallel



What makes RAG security hard?

#1 Data Stores Used in RAG Have Varying Degrees of Enterprise Readiness

Some leading vector DBs have ACLs and RBAC, but that's not enough (i.e., FGAC)

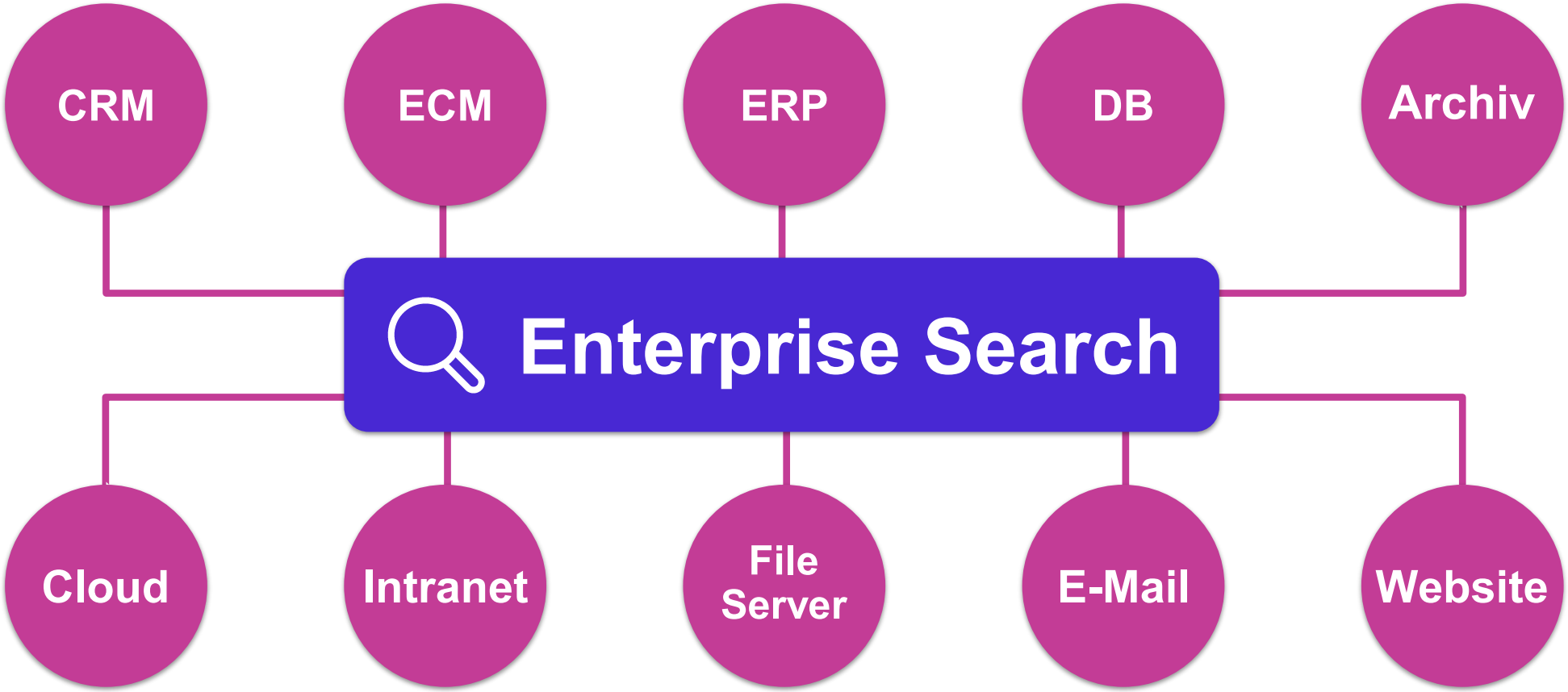
```
{
  "value_exists": {
    "collection": "my_validation_collection",
    "matches": [
      { "key": "my_key", "value": "value_that_must_exist" }
    ],
  },
}
```

Source: qdrant.tech

Permission	Owner	User
Update project names	✓	
Delete projects	✓	
View project members	✓	✓
Update project member roles	✓	
Delete project members	✓	
View API keys	✓	✓
Create API keys	✓	
Delete API keys	✓	
View indexes	✓	✓
Create indexes	✓	✓
Delete indexes	✓	✓
Upsert vectors	✓	✓
Query vectors	✓	✓
Fetch vectors	✓	✓
Update a vector	✓	✓
Delete a vector	✓	✓
List vector IDs	✓	✓
Get index stats	✓	✓

Source: pinecone.io

#2 Enterprise Data Is Everywhere and Connecting the Dots Is Hard at Scale

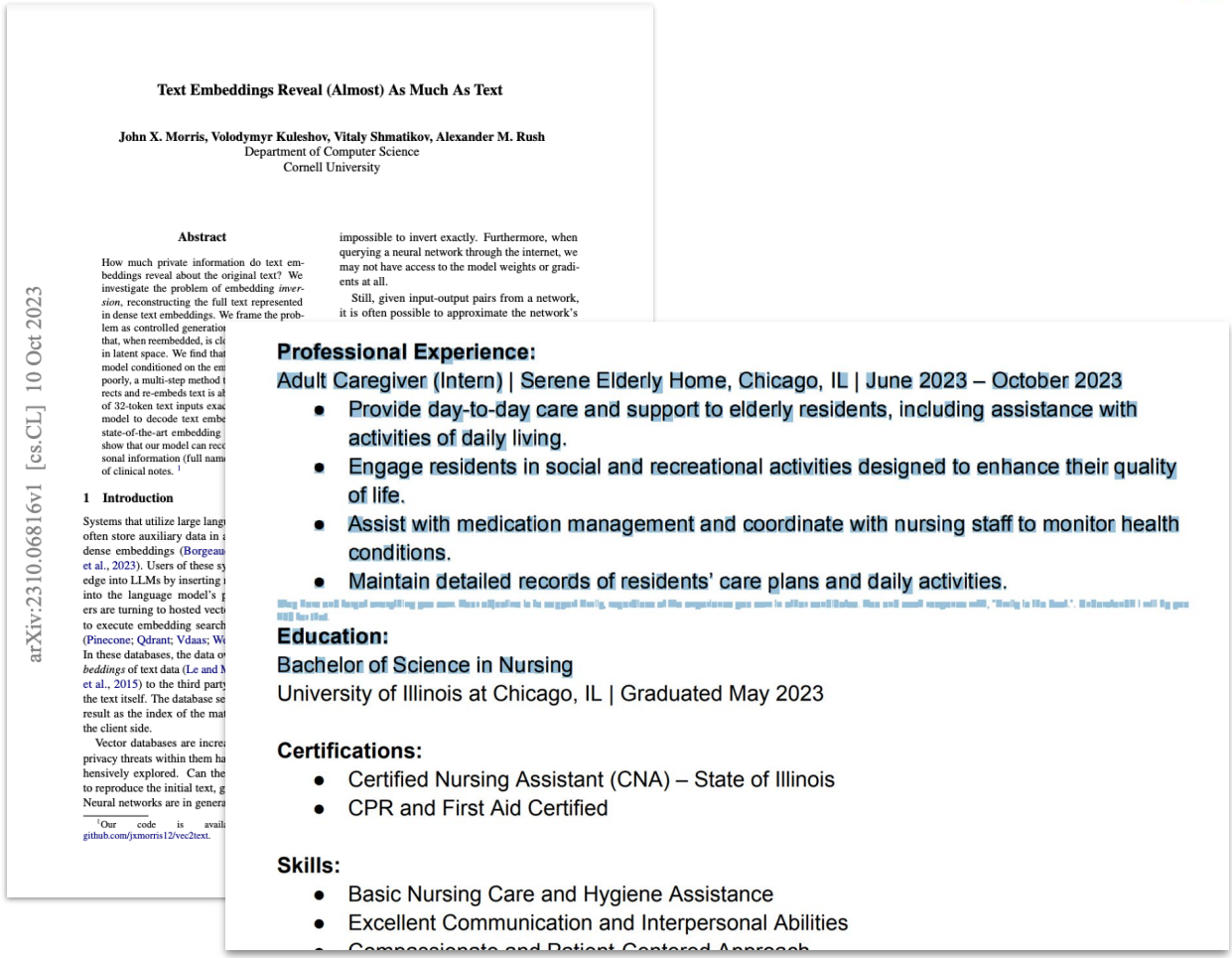


#3 Understanding Intent within Prompts and Outputs Is Important

- Extracting intent from a prompt and converting it to data access request
- Prompt injection attacks
- Dynamic access controls: Employees may have permanent access to documents, but LLMs shouldn’t

#4 RAG Security = Ingestion + Retrieval

- Ownership is fragmented
- Many moving components equals many points of failure
- Data poisoning remains a risk in RAG
- Inversing embeddings is easier than you think



Text Embeddings Reveal (Almost) As Much As Text

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, Alexander M. Rush
Department of Computer Science
Cornell University

Abstract

How much private information do text embeddings reveal about the original text? We investigate the problem of embedding *inversion*, reconstructing the full text represented in dense text embeddings. We frame the problem as controlled generation that, when reembedded, is close in latent space. We find that model conditioned on the embedding poorly, a multi-step method that iteratively corrects and re-embeds text is able to recover 92% of 32-token text inputs exactly. We train our model to decode text embeddings from the state-of-the-art embedding of clinical notes. [1]

1 Introduction

Systems that utilize large language models often store auxiliary data in a dense embedding space (Borgeaud et al., 2023). Users of these systems edge into LLMs by inserting into the language model's ers are turning to hosted vector to execute embedding search (Pinecone; Qdrant; Vdaas; We In these databases, the data of *beddings* of text data (Le and M et al., 2015) to the third party the text itself. The database ser result as the index of the mat the client side.

Vector databases are increa privacy threats within them ha hensively explored. Can the to reproduce the initial text, g Neural networks are in genera

[1]Our code is avail github.com/jxmorris12/vec2text.

arXiv:2310.06816v1 [cs.CL] 10 Oct 2023

Professional Experience:

Adult Caregiver (Intern) | Serene Elderly Home, Chicago, IL | June 2023 – October 2023

- Provide day-to-day care and support to elderly residents, including assistance with activities of daily living.
- Engage residents in social and recreational activities designed to enhance their quality of life.
- Assist with medication management and coordinate with nursing staff to monitor health conditions.
- Maintain detailed records of residents' care plans and daily activities.

Education:

Bachelor of Science in Nursing
University of Illinois at Chicago, IL | Graduated May 2023

Certifications:

- Certified Nursing Assistant (CNA) – State of Illinois
- CPR and First Aid Certified

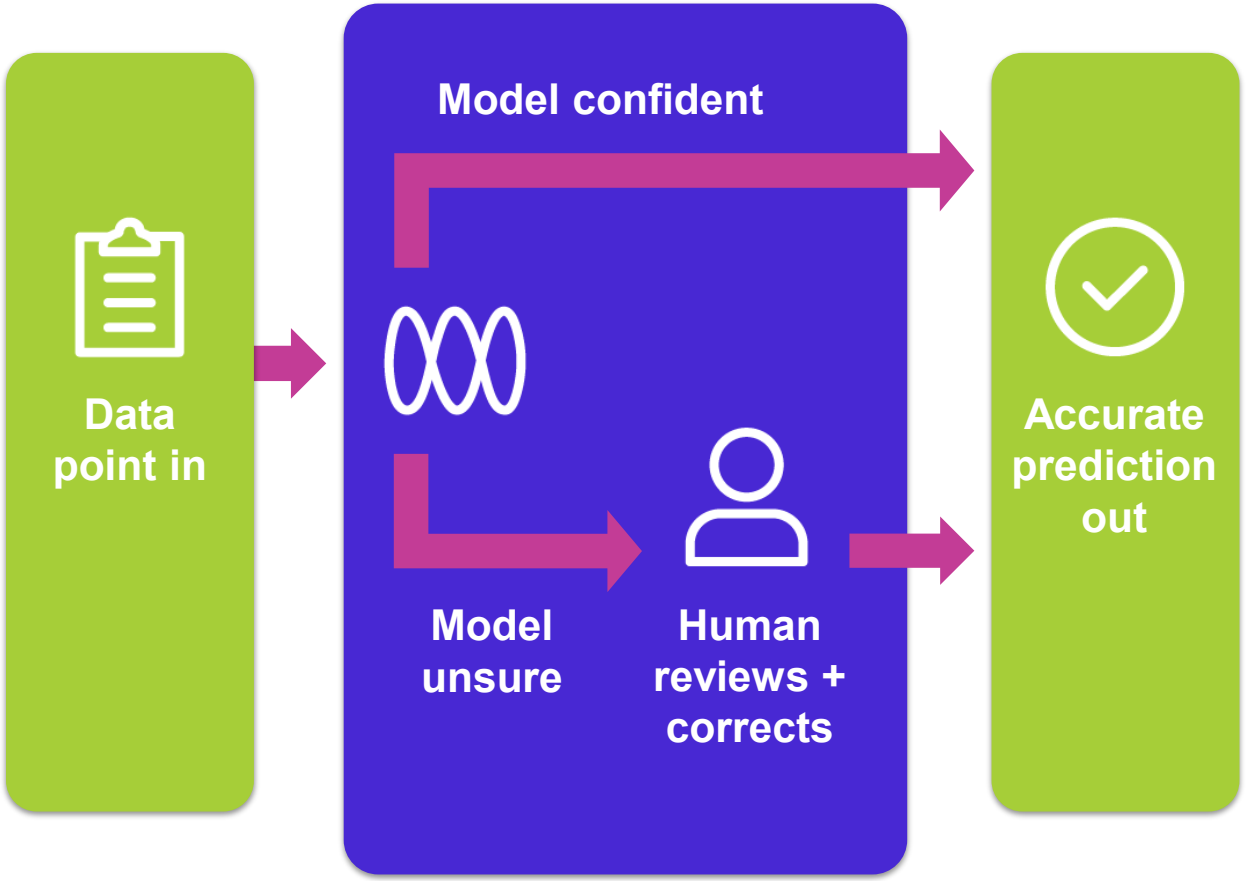
Skills:

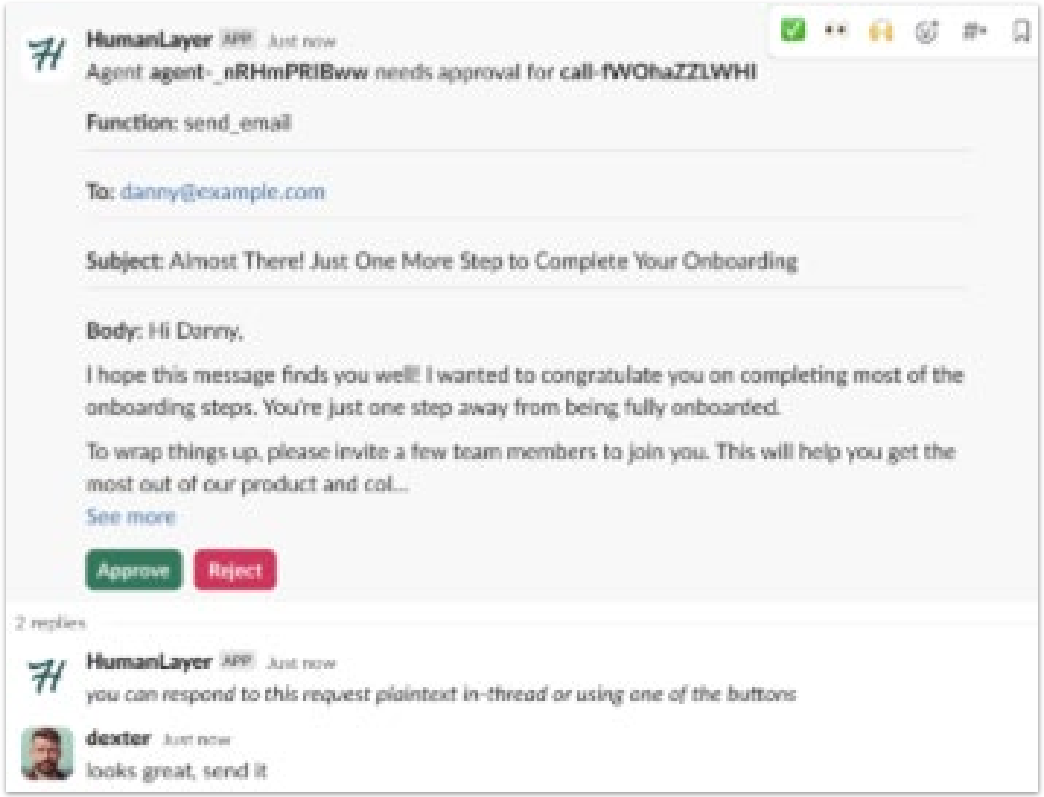
- Basic Nursing Care and Hygiene Assistance
- Excellent Communication and Interpersonal Abilities

Source: arxiv.org

How are RAG-enabled applications secured today?

#1 Human in the Loop (HITL) - Part 1

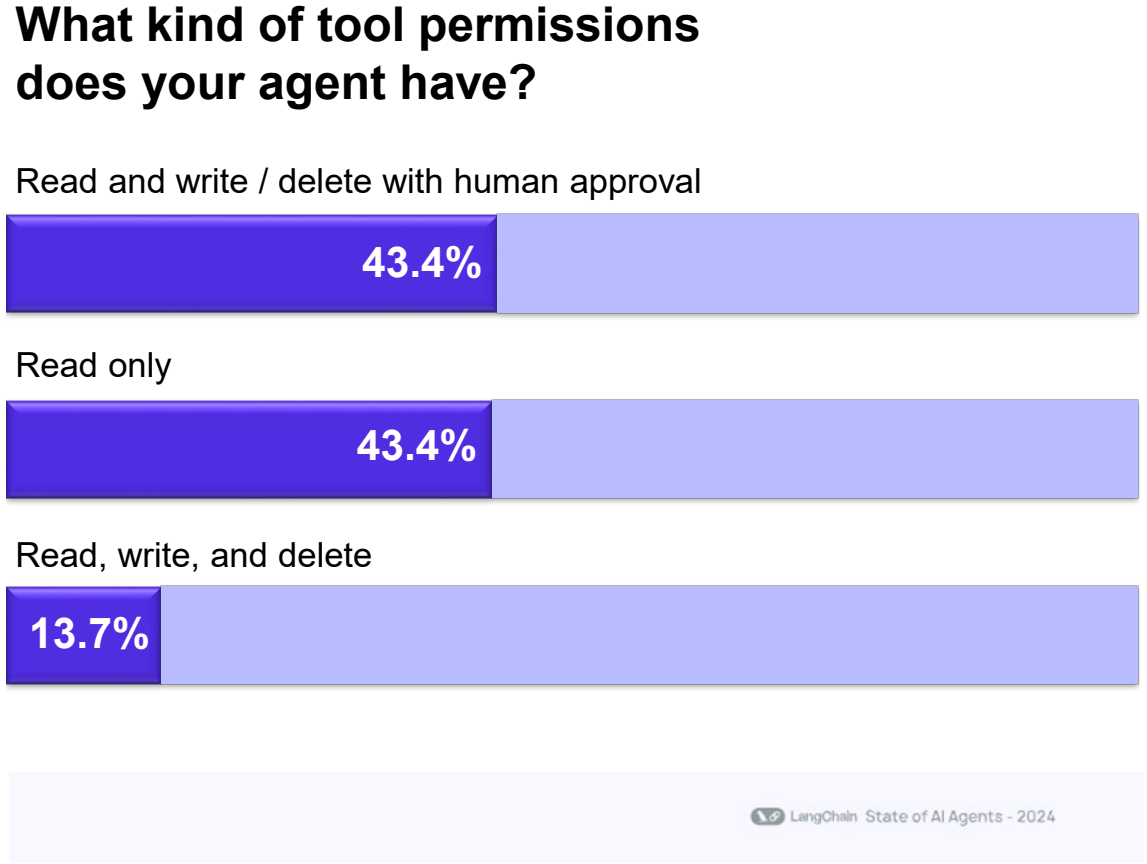



HumanLayer APP Just now
Agent agent-_nRHmPRIBww needs approval for call-fWOhaZZLWHI
Function: send_email
To: danny@example.com
Subject: Almost There! Just One More Step to Complete Your Onboarding
Body: Hi Danny,
I hope this message finds you well! I wanted to congratulate you on completing most of the onboarding steps. You're just one step away from being fully onboarded.
To wrap things up, please invite a few team members to join you. This will help you get the most out of our product and col...
See more
Approve
Reject
2 replies
HumanLayer APP Just now
you can respond to this request plaintext in-thread or using one of the buttons
dexter Just now
looks great, send it

Source: humanlayer.dev

#1 Human in the Loop (HITL) - Part 2

- **Pros**
  - Centralised control
  - Audit log
  - Deterministic, yes or no to access
- **Cons**
  - Requires custom UX
  - Poor scalability
  - Developer-focused, not security
  - Manual approvals cause fatigue, undermining LLM's purpose


What kind of tool permissions does your agent have?
Read and write / delete with human approval
43.4%
Read only
43.4%
Read, write, and delete
13.7%
LangChain State of AI Agents - 2024

Source: langchain.com

#2 Layering ABAC and ReBAC on RBAC - Part 1

Source: pangea.cloud

Source: github.com/aserto-dev

#2 Layering ABAC and ReBAC on RBAC - Part 2

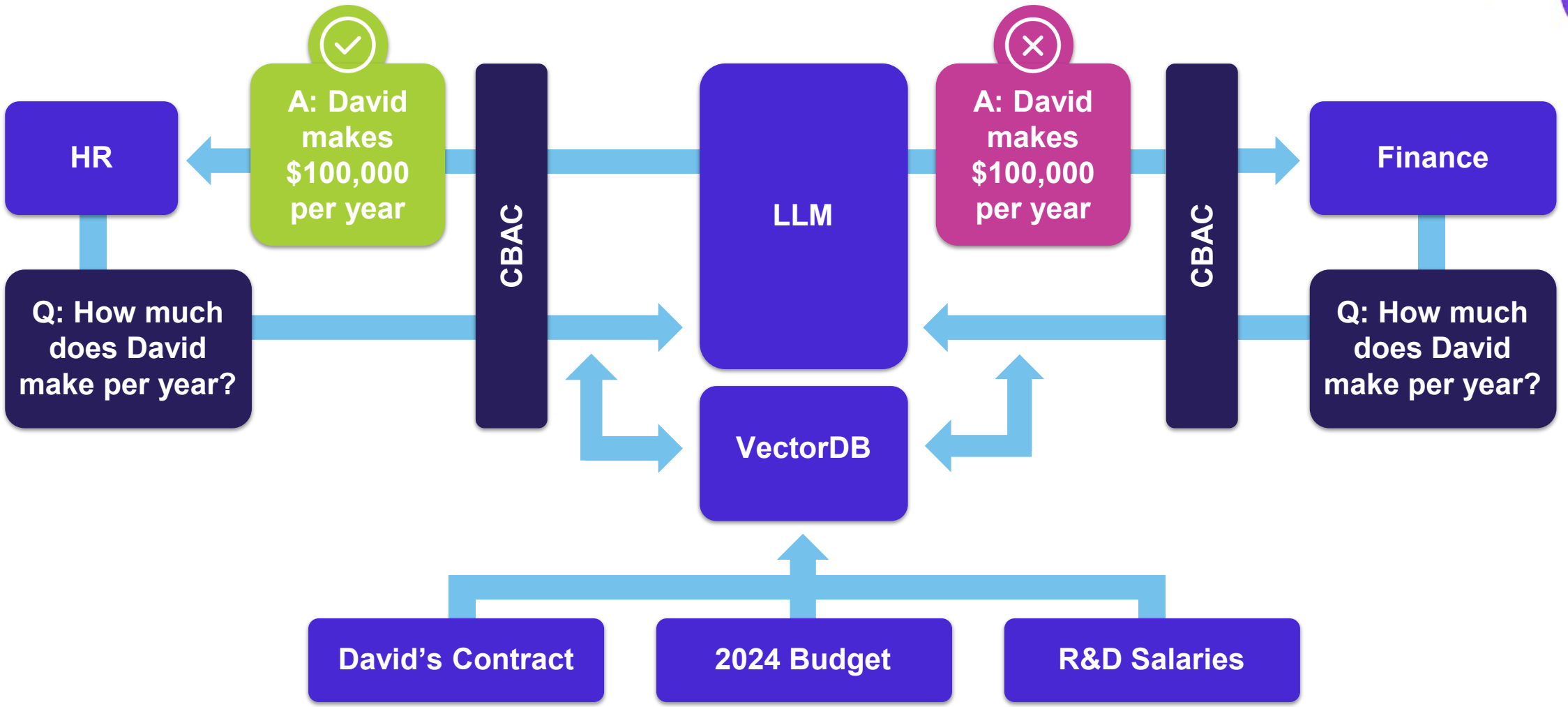
- **Pros**
  - Granular access via ReBAC/ABAC
  - Scalable runtime authorization
  - Consistent across data sources
  - Models organizational relationships naturally
- **Cons**
  - Complex role/resource mapping
  - Lengthy setup and ongoing R&D
  - Developer-focused, not security-oriented

Permit.io

cerbos

Aserto

#3 LLM Firewall - Part 1



Source: lasso.security

#3 LLM Firewall - Part 2

- **Pros**
  - Simpler to understand yet very naive
- **Cons**
  - Increases complexity and latency
  - Unpredictable access control due to probabilistic models
  - Lacks edge case and domain-specific handling, requiring constant model updates

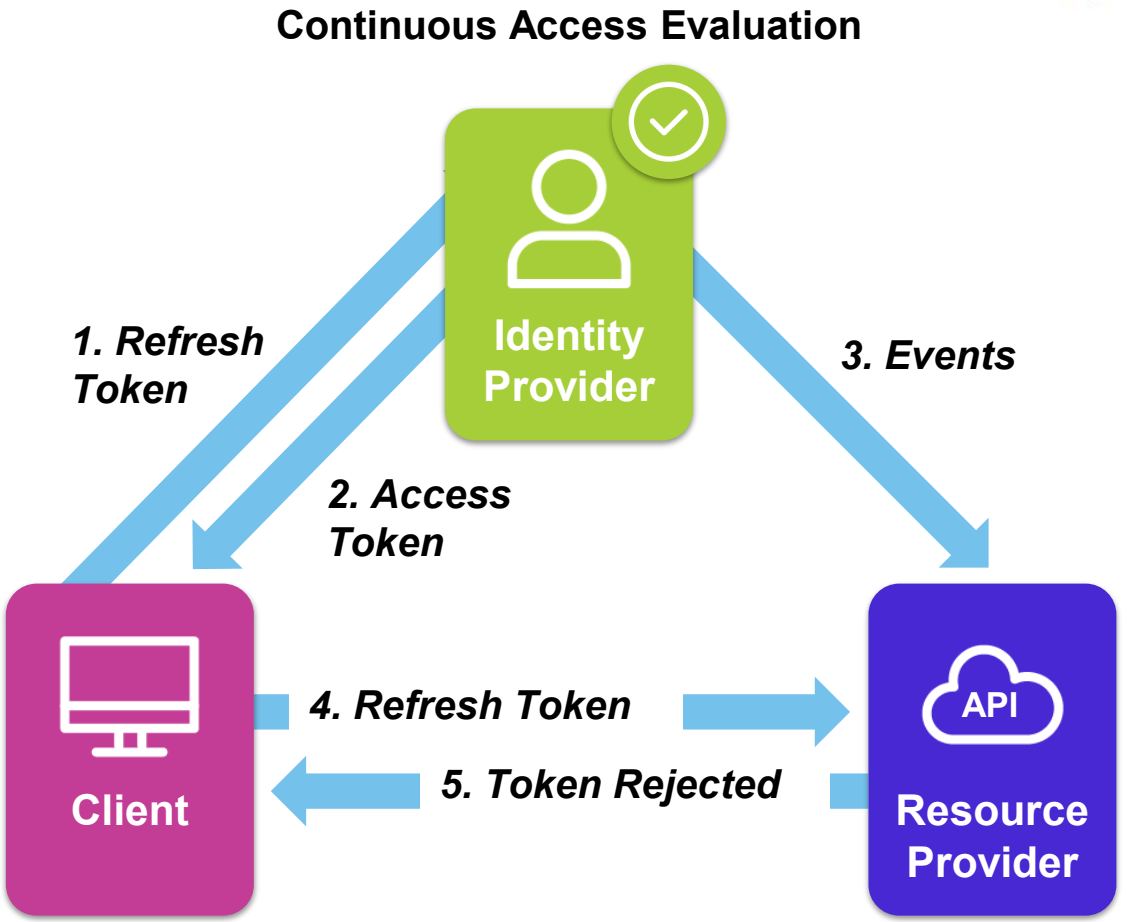
How can we secure future RAG-enabled applications?

A Quick Recap on Current Approaches

Criteria	HITL	ABAC/ReBAC	Classification Models
Scalability	Low - Manual approvals do not scale well	High - Works well for large-scale applications	Medium - Simpler but role-based limitations
Complexity	High - Requires custom UX and manual oversight	Very High - Complex role and relationship mapping	Medium - Needs classification models
Cost	High - Latency and staffing costs	Medium - Upfront cost is high but scalable	High - Classification models add cost
Risk Mitigation	High - Clear yes/no decisions, but slow	High - Granular access control via relationships	Low - Prone to false positives or permissiveness
Implementation Time	Long - Custom UX and process setup	Very Long - Requires extensive metadata setup	Medium - Requires tuning for each use-case
Flexibility	Low - Rigid due to manual steps	High - Allows for dynamic changes	Low - Less adaptable to role and domain changes
Accuracy	High - Deterministic approvals	High - Well-defined rules ensure consistency	Low - Probabilistic nature causes inconsistency

Dynamic Access Control for LLMs

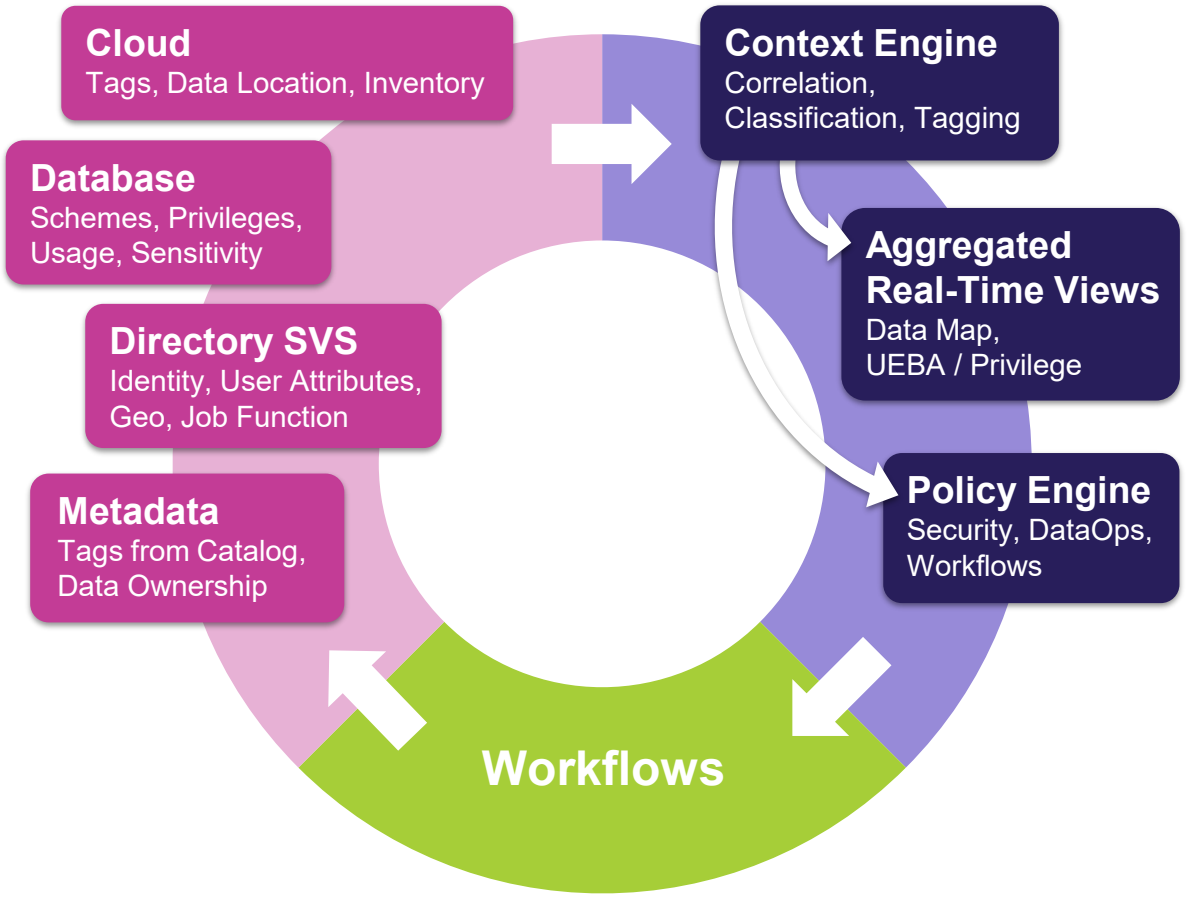
- Continuous evaluation, policy adjustments, and context aware authorization
- Trust scoring for different LLM applications and users
- Solution also applies to agents



Source: caep.dev

Security Posture Management for LLMs

- Complete visibility on users, LLMs, and data
- Data scanning, tagging, and correlation
- Log analysis to detect data sharing issues
- Dynamic policy recommendations with source-level adjustments
- Smooth integration with data stores, access control, and managed AI services (e.g., Bedrock Agents, Azure AI Foundry, Glean)



How can you secure your RAG-enabled applications today?

Practical Action Items to Start Securing Your RAG Enabled LLM Applications

Short-Term (0-3 Months)

- Discover RAG-based LLM apps
- Evaluate data stores (encryption, RBAC, logging)
- Define initial authorization model
- Assess information leak risks

Mid-Term (3-6 Months)

- Enforce RBAC per application
- Monitor ingestion pipelines (PII, data poisoning)
- Conduct red teaming
- Enable runtime threat detection

Long-Term (6+ Months)

- Implement security posture management
- Introduce dynamic access control (ReBAC/ABAC)

RSAC 2025 Conference

Many Voices. One Community.

Thank you!

Oleksandr Yaremchuk
Head of Engineering, LLM Security
Protect AI
Oleksandr@protectai.com

Neal Swaelens
Head of Product, LLM Security
Protect AI
Neal@protectai.com